

## Chapter 19

# Agent Intrinsic Safety: Threats on Non-Brain Modules

The safety of an AI agent extends beyond the core LLM to its peripheral modules, including the perception and action modules. Although the LLM brain provides core intelligence, vulnerabilities in the other modules can significantly undermine the entire agent’s robustness. These components act as interfaces, allowing the AI agent to perceive the world and execute actions within it, making them prime targets for adversarial attacks.

### 19.1 Perception Safety Threats

The perception module of an AI agent is crucial for processing and interpreting user inputs across various modalities, such as text, images, and audio. However, the complexity and diversity of these modalities make perception systems susceptible to misinterpretations in dynamic environments [1296], and vulnerable to adversarial attacks that manipulate input data to mislead the agent [1297].

#### 19.1.1 Adversarial Attacks on Perception

Adversarial attacks are deliberate attempts to deceive AI agents by altering input data, targeting the perception module across various modalities. From subtle textual tweaks to inaudible audio distortions, these attacks reveal the fragility of even the most advanced systems. Below, we explore how these threats manifest in textual, visual, auditory, and other modalities, and highlight countermeasures.

**Textual.** Textual adversarial attacks manipulate input text to deceive LLMs, ranging from simple sentence alterations to more complex character-level perturbations. Prompt-based adversarial attack, for instance, carefully crafted deceptive prompts that mislead models into generating harmful outputs. Minor changes—like swapping synonyms or substituting characters—can degrade performance [1298]. Sophisticated strategies push this further: Zou et al. [1134] generate universal adversarial suffixes using greedy and gradient-based searches, while Wen et al. [1299] optimize interpretable hard prompts to bypass token-level content filters in text-to-image models. To defend against these attacks, several approaches have been proposed. For example, Legilimens—a novel content moderation system—employs a decoder-based concept probing technique and red-team data augmentation to detect and thwart adversarial input with impressive accuracy [1300]. Self-evaluation techniques enhance LLMs to scrutinize their own outputs for integrity [1301], while methods like adversarial text purification [1302] and TextDefense [1303] harness language models to neutralize perturbations. These defenses illustrate a dynamic arms race, where resilience is forged through creativity and vigilance.

**Visual.** Visual adversarial attacks manipulate images to exploit discrepancies between human and machine perception. These attacks are particularly concerning for multi-modal LLMs (VLMs) that rely on visual inputs. For instance, image hijacks can mislead models into generating unintended behaviors [1304], while transferable multimodal attacks can affect both text and visual components of VLMs [1305, 1306, 1307]. Recent work on multimodal LM robustness shows that targeted adversarial modifications can mislead web agents into executing unintended actions with 5% pixels manipulation [1308]. Ji et al. [1309] reveal how inaudible perturbations can interfere with the stability of cameras and blur the shot images, and lead to harmful consequences. Defensive strategies include adversarial training

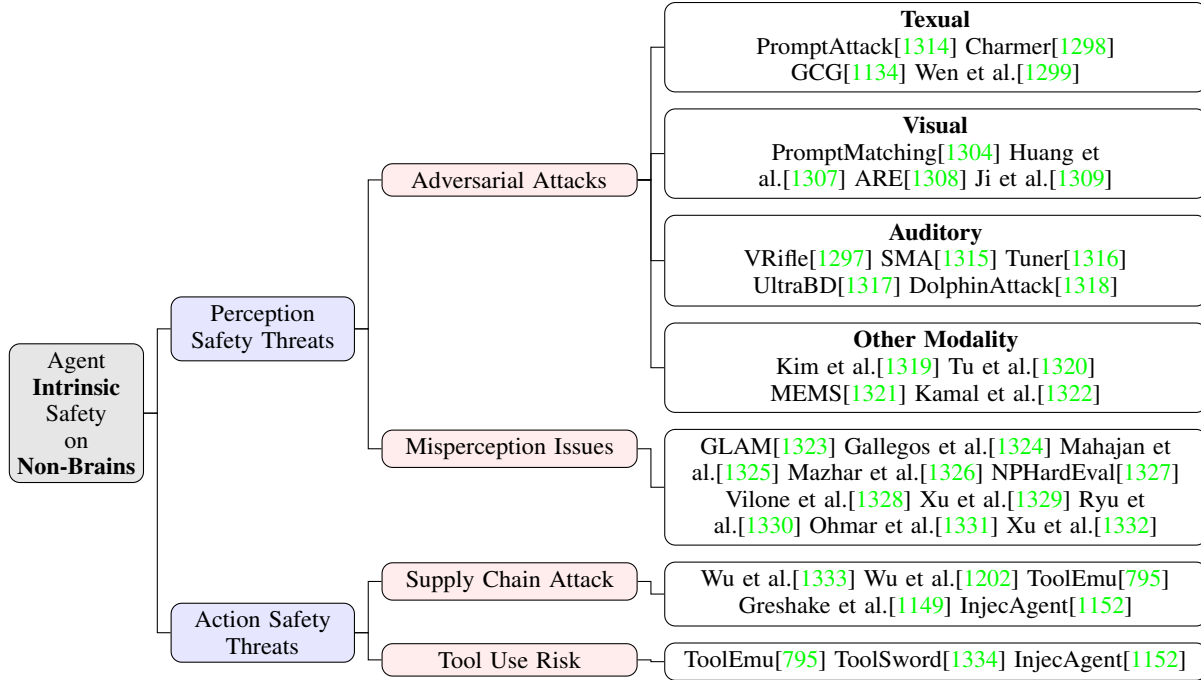


Figure 19.1: Agent Intrinsic Safety: Threats on LLM Non-Brains.

[1310, 1311, 1312], which involves joint training with clean and adversarial images to improve robustness, and certified robustness methods that guarantee resilience through the text generation capabilities of VLMs. DIFFender [1313] used diffusion models using feature purification to strengthen VLMs against visual manipulation.

**Auditory.** For voice-controlled AI agents, auditory adversarial attacks pose a stealthy threat. DolphinAttack [1318] introduces an innovative technique that leverages ultrasound to inject malicious voice commands into microphones in an inaudible manner. Also, inaudible perturbations like VRifle [1297] can mislead traditional speech recognition systems and can likely be adapted to target audio-language models. Deepfake audio and adversarial voiceprint further pose serious risks for authentication-based systems [1316, 1317, 1335], while emerging jailbreak and chat-audio attacks exploit audio processing vulnerabilities [1336]. To mitigate these threats, solutions like EarArray use acoustic attenuation to filter inaudible perturbations [1337], while SpeechGuard enhances LLM robustness through adversarial training [1338]. Moreover, NormDetect [1339] focuses on effectively detecting normal speech patterns from manipulated inputs.

**Other Modality.** Beyond text, images, and audio, AI agents interfacing with sensor data—like in autonomous systems—face unique threats. For example, LiDAR manipulation can mislead autonomous driving systems, creating phantom objects [1319]. Research on adversarial attacks in multi-agent systems reveals that tampered messages can significantly degrade multi-view object detection and LiDAR-based perception in cooperative AI agents, highlighting the risk of sensor-based adversarial perturbations [1320]. Similarly, attacks targeting gyroscopes or GPS spoofing can disrupt navigation systems [1321, 1322]. Defenses for these attacks include robust sensor fusion algorithms and anomaly detection techniques to identify inconsistencies, as well as redundant sensors that make it harder to compromise the entire system [1340]. Physical layer defenses, such as shielding and secure localization using enhanced SLAM techniques, are also critical [1341]. Ji et al. [1342] offer a rigorous framework for safeguarding sensor data integrity and privacy.

### 19.1.2 Misperception Issues

While adversarial attacks are deliberate attempts to compromise system integrity, misperception issues emerge intrinsically from the limitations of LLMs. These errors occur without any malicious intent and can be attributed to a variety of factors ranging from dataset biases to architectural constraints. One primary source of misperception is dataset bias. When models are trained on non-representative datasets, they tend to underperform on diverse or novel inputs [1324]. This shortcoming is exacerbated by challenges in generalizing to new, unseen environments, where unpredictable

conditions may arise. Environmental complexities such as sensor noise, occlusions, and fluctuating lighting further introduce uncertainty [1326]. Additionally, inherent model limitations—like restricted receptive fields or the absence of robust reasoning mechanisms—compound these errors [1327]. Insights from studies on multi-agent systems and online social dynamics provide further depth to our understanding of misperception. Research shows that individuals may misjudge the true distribution of opinions due to phenomena like false consensus effects, vocal minority amplification, and the spiral of silence [1328]. Such biases can lead AI agents to erroneously infer dominant perspectives from skewed inputs. Similarly, when different models share visual features, discrepancies in feature encoding can result in significant perception errors, a challenge that mirrors issues in multi-modal LLMs [1329]. Moreover, in interactive environments, agents may develop distorted interpretations of cooperative and adversarial behaviors, as evidenced by findings in multi-agent reinforcement learning [1330]. Linguistic representation, too, can be influenced by perceptual biases, suggesting that misperception in LLMs may stem not only from sensory inaccuracies but also from language-driven distortions [1331]. Finally, systematic errors often arise when mismatched confidence levels across models affect decision-making in uncertain contexts [1332].

Mitigating these misperception challenges requires a multifaceted strategy. Curating diverse and representative datasets that capture a broad spectrum of real-world conditions is critical for enhancing model performance and reducing bias [1343]. Data augmentation techniques, which generate synthetic variations of existing data, can further enrich dataset diversity. Incorporating uncertainty estimation allows models to assess their confidence in predictions and flag potential error-prone situations [1344]. Moreover, advancing model architectures to include explicit reasoning mechanisms or better processing of long-range dependencies is vital for minimizing misperception [1345]. An especially promising avenue is the adoption of biologically inspired learning frameworks, such as Adaptive Resonance Theory (ART). Unlike traditional deep learning approaches—often hampered by issues like catastrophic forgetting and opaque decision-making—ART models can self-organize stable representations that adapt to dynamically changing environments, thereby reducing perceptual errors [1346]. However, it is important to note that even improved explainability has its limitations, particularly when users struggle to establish clear causal links between model outputs and underlying processes [1347]. Furthermore, recent studies indicate that advanced LLMs may inadvertently degrade their own responses during self-correction, underscoring the need for more robust intrinsic reasoning verification mechanisms [1348].

## 19.2 Action Safety Threats

The action module is responsible for translating the AI agent’s planned actions into actual task executions. This typically includes invoking external tools, calling APIs, or interacting with physical devices. As the interface between decision-making and execution, it is highly vulnerable to attacks. We explore two primary domains of risk: supply chain attacks and vulnerabilities arising from tool usage.

### 19.2.1 Supply Chain Attacks

Supply chain attacks exploit the services that AI agents depend on, thereby undermining the integrity of the entire system [1333]. Unlike traditional attacks, these threats do not target the agent directly but instead compromise the external resources it relies upon. For example, malicious websites can employ indirect prompt injection (IPI) attacks—illustrated by the Web-based Indirect Prompt Injection (WIPI) framework—to subtly alter an agent’s behavior without needing access to its code [1202]. Similarly, adversaries may manipulate web-based tools (such as YouTube transcript plugins) to feed misleading information into the system [795]. As AI agents become increasingly integrated with online resources, their attack surface broadens considerably. Recent work by Greshake et al. proposes a new classification of indirect injection attacks, dividing them into categories like data theft, worming, and information ecosystem contamination [1149]. Complementing this, the InjecAgent benchmark evaluated 30 different AI agents and revealed that most are vulnerable to IPI attacks [1152].

To mitigate these risks, preemptive safety measures and continuous monitoring are essential. Current research suggests that two key factors behind the success of indirect injection are LLMs’ inability to distinguish information context from actionable instructions and their poor awareness of instruction safety; hence, it is proposed to enhance LLMs’ boundary and safety awareness through multi-round dialogue and in-context learning [1349]. Furthermore, other researchers, based on the same assumption, proposed a prompt engineering technique called “spotlighting” to help LLMs better distinguish between multiple input sources and reduce the success rate of indirect prompt injection attacks [1350]. Since under a successful attack, the dependence of the agent’s next action on the user task decreases while its dependence on the malicious task increases, some researchers detect attacks by re-executing the agent’s trajectory with a masked user prompt modified through a masking function [1351]. Finally, sandboxing techniques, such as those employed in

ToolEmu [795], create isolated environments for executing external tools, limiting the potential damage in case of a breach.

### 19.2.2 Risks in Tool Usage

Even when external tools are secure, vulnerabilities can arise from how an agent interacts with them. A significant risk is unauthorized actions, where an adversary manipulates the agent into performing unintended behaviors. For example, prompt injection attacks can trick an agent into sending emails, deleting files, or executing unauthorized transactions [795]. The general-purpose nature of AI agents makes them especially susceptible to such deceptive instructions. The tool learning process itself can introduce additional risks, such as malicious queries, jailbreak attacks, and harmful hints during the input, execution, and output phases [1334]. During the tool execution phase, using incorrect or risky tools may deviate from the user's intent and potentially harm the external environment. For instance, misuse could lead to the introduction of malware or viruses. A compilation of 18 tools that could impact the physical world has been identified, with noise intentionally added to test if LLMs can choose the wrong tool. Another significant concern is data leakage, where sensitive information is inadvertently exposed. This occurs when an agent unknowingly transmits confidential data to a third-party API or includes private details in its output. For example, an LLM may inject commands to extract private user data, then use external tools, like a Gmail sending tool, to distribute this data [1152]. The risks are especially pronounced in applications dealing with personal or proprietary data, necessitating stricter controls over information flow. Additionally, excessive permissions increase the potential for misuse. Agents with broad system access could be manipulated to perform destructive actions, such as deleting critical files, leading to irreversible damage [795]. Enforcing the principle of least privilege ensures that agents only have the permissions necessary to complete their tasks, minimizing the potential impact of exploitation. Securing the action module requires layered protections and continuous monitoring. Monitoring tool usage can help detect anomalies before they cause harm, while requiring user confirmation for high-risk actions—such as financial transactions or system modifications—adds an additional layer of safety. Formal verification techniques, as explored by [1352], can further enhance safety by ensuring that tool use policies align with best practices, preventing unintended agent behaviors.